Version 16 December, 2019. This article has been accepted for publication by Edinburgh University Press in *Corpora* (https://www.euppublishing.com/loi/cor). It is scheduled to come out in issue 16.2 in 2021.

Timo Korkiakangas, University of Helsinki

Late Latin Charter Treebank: contents and annotation

Abstract

This paper describes the construction and annotation of the Late Latin Charter Treebank, a set of three dependency treebanks (LLCT1, LLCT2, and LLCT3) which contain together 1,261 Early Medieval Latin documentary texts (i.e., original charters) written in Italy between AD 714 and 1000 (c. 594,000 tokens). The paper focuses on issues which a linguistically or philologically inclined user of LLCT needs to know: the criteria on which the charters were selected, the special characteristics of the annotation types utilized and the geographical and chronological distribution of the data. In addition to normal queries on forms, lemmas, morphology and syntax, complex philological research settings are enabled by the textual annotation layer of LLCT, which indicates abbreviated and damaged words, as well as the formulaic and non-formulaic passages of each charter.

1. Motivation

Based on frequent personal requests as well as the high volume of downloads of LLCT1 at Zenodo, the research community seems to be in need of a dependency treebank of late, non-standard Latin, which permits comparison with Classical Latin treebanks, such as the Latin Dependency Treebanks (LDT) or the PROIEL treebanks, as well as with the Index Thomisticus Treebank (IT-TB) of Medieval Latin.[1] Therefore, even the recently finished LLCT2 was made available at Zenodo and will be published later in the Universal Dependencies. Consequently, it is necessary to provide an up-to-date description of the entire LLCT treebank. The last part, LLCT3, will be completed and published by 2021. Some aspects of the annotation and contents of LLCT1 were discussed in Korkiakangas and Passarotti (2011), Korkiakangas and Lassila (2018), and Korkiakangas (2016), but a synopsis of the data has thus far been lacking.

2. Charters as a linguistic resource

LLCT consists of *charters*, legal documents of typically 200–1,000 words in length, written by quill on a single sheet of parchment. They survive in the thousands in Italy from the Early Middle Ages.

[1] LDT: https://perseusdl.github.io/treebank_data/, PROIEL: https://proiel.github.io, IT-TB: https://itreebank.marginalia.it. For an overview of computational resources and tools for Latin, see McGillivray, 2014 (esp. §2).

Charters are written by professional or unprofessional writers to record private transactions, such as the sale, exchange and lease of property, or as semi-public memoranda of trials (Pratesi, 1979).

Classical Latin is the form of Latin that the writers of the late Roman Republic and the Roman Empire established as a standard. Its spelling and grammar were still considered the only viable option by the literary authors of the Late Antiquity and even afterwards. The Latin of charters, on the other hand, is often described as 'non-standard' with respect to Classical Latin: there is much variation both from writer to writer and from Classical Latin in terms of spelling, lexicon, morphology and syntax. Consequently, charters constitute a useful resource for understanding the evolution of Latin in the crucial period of transition between Latin and Romance (Sornicola, 2017). Charters almost always convey the date, the name of the scribe, and the writing place, enabling the modern-day scholar to make diachronic, sociolinguistic and diatopic comparisons. In addition, the charters are originals, so their wording tends to reflect the language use of the date indicated, contrary to literary texts, which are always mediated by multiple manuscript generations.

However, several limitations exist in the use of charters as a linguistic resource. First, the documents are formulaic—i.e., they rehearse prefabricated *formulae* to guarantee legal validity. Although wide variation in the formulae proves that charters were not simply copied from model document collections, each charter nevertheless repeats the conventional wordings relative to its document type, which the writer had memorized. Consequently, LLCT has a peculiar type-token ratio (e.g., only 0.04 in LLCT2). This entails that some linguistic features are not attested at all, while others are over-represented. On the other hand, each document also has one or more non-formulaic—so-called *free*—parts, where the details related to that specific legal case are given. The non-formulaic parts better reflect the spoken language, because they cannot rely on prefabricated formulae (Sabatini, 1965). Thus, although not everything is formulaic in the documents, they still only offer a narrow panorama of the linguistic landscape. With their formulaic nature, however, charters offer a kind of parallel to elicited texts, which are typically used in language acquisition studies and which focus on certain limited linguistic phenomena under scrutiny.

Another limitation of charters is their relation to the spoken language. As formulae draw on centuries-old legal Latin tradition, the linguistic features that differ from Classical Latin grammar do not necessarily reflect developments of the spoken Latin of the time, but may be misunderstandings or archaisms of individual scribes or scribal schools. Because the distance between the written and spoken codes expanded greatly along with language evolution, Early Medieval writers learnt their Latin practically as a second language (Korkiakangas 2018).

Note that, by calling the language of LLCT 'Late' instead of 'Medieval' Latin, it is postulated that the writers did not conceptually separate the language they wrote from the spoken language of the time, contrary to what was the case later with Medieval Latin (cf. Wright, 2016). Despite its formulaic nature, the Latin of LLCT reflects the uninterrupted evolution of the spoken Latin. Latin was still the only idiom available with an established written form, though the first known reliably datable vernacular text excerpts start appearing in the 9th and 10th centuries (Frank-Job & Selig, 2016).

3. Building LLCT

LLCT consists of three parts (LLCT1, LLCT2 and LLCT3), built independently one after another beginning from 2010. Table 1 presents the central features of the LLCT parts. LLCT3 is currently

being annotated, with an expected completion time of 2021—hence, the approximations in the data for LLCT3. The effective word count excludes punctuation and words with fragmentary or abbreviated inflectional endings.

Table 1. LLCT, background information.

Treebank part	Constructed	Charters	Tokens	Effective words	Location	Years
LLCT1	2010-14	519	225,834	198,696	Tuscany	714–869
LLCT2	2016-18	521	257,918	220,797	Tuscany	774–897
LLCT3	2019-	~ 221	~ 110,400	~ 94,000	Tuscany, North and South Italy	721–1000

The whole LLCT is based on the technical choices and annotation styles made available in the Perseus Digital Library Project by 2010. The dependency relations of LLCT1 were drawn using the graphical Alpheios interface, which was linked to the morphological table editor of the Perseus annotation environment, now replaced by the Arethusa annotation tool.[2] In the table editor, the Morpheus morphological analyser proposed one or more lemmas and morphological analyses for each form. Since the spelling and grammar of LLCT are non-standard, the annotation environment only directly provided a correct analysis of 20–40% of the forms, depending on charter. As for the remaining cases, a lower-ranking analysis had to be picked from a drop-down menu, or, if it was lacking completely, it was added manually to a repository of user-saved analyses. LLCT2 was then annotated using machine learning on LLCT1 and corrected manually.

Each LLCT charter is provided with metadata attributes indicating the date, scribe, place of writing, and document type. As for graphical choices, the letter *j* was converted to *i* in the entire LLCT, while prevocalic *u* is indicated as either *u* or *v*, depending on the original edition. In LLCT1, the traditional Latin convention is followed, with forms and lemmas that indicate months and calendar terms, such as Kalends, along with normal proper names, capitalized. In LLCT2 and LLCT3, on the other hand, only proper-name lemmas are capitalized. Punctuation in all cases follows that of the printed-source editions (see section 5) but was changed where necessary. With the 19th-century editions punctuation was changed more radically to comply with modern-day requirements.

4. The LLCT-specific annotation practices

The syntactic annotation of LLCT follows dependency grammar and is mostly compatible with the style presented in the *Guidelines for the Syntactic Annotation of Latin Treebanks* (version 1.3, Bamman & al., 2007; henceforth *Guidelines*), while lemmatization was simply done in imitation of that of the LDT treebanks available in 2010. The then-existing LDT annotations also served as the model for the morphological annotation of standard Latin forms, but, for Late Latin non-standard forms, an additional set of rules was designed (Korkiakangas and Passarotti, 2011). Some modifications had to be introduced to the syntactic *Guidelines* as well due to the special nature of charter Latin. This section presents the annotation choices that depart from syntactic annotation as described in the *Guidelines* or from the morphological annotation applied in LDT.

[2] Arethusa: https://www.perseids.org/apps/treebank

The *Guidelines* propose twenty syntactic tags to be used within the framework of dependency grammar, where each word is linked as a modifier to its immediate head. The direct models are the Prague Dependency Treebank (PDT)[3] and Latin grammar as codified in Pinkster 1990. In many respects, the *Guidelines* represent the traditional description of Latin pedagogical grammar (*Guidelines*: 3). The highest node of each dependency tree is a finite verb (tagged PRED), attached to the sentence root, except when it is coordinated. Coordination is represented by a superordinate COORD node and a _CO extension of the coordinated nodes. Other 'bridge structures' include preposition (AuxP) and conjunction (AuxC). They function as 'bridges' between their complements and their own heads (*Guidelines*: 26–29). In the typical Latin accusativus cum infinitivo construction, the accusative-form subject (SBJ) depends on the infinitive, which is marked as the object (OBJ) (*Guidelines*: 6, 42). Relative clauses are annotated based on their syntactic function in the sentence: for example, the finite verb of an attributive relative clause depends on its antecedent and receives the tag ATR (*Guidelines*: 37–38). In LLCT1, punctuation (AuxX, AuxG, AuxK), non-nominal sentence adverbials (AuxY) and emphasizing particles (AuxZ) (*Guidelines*: 30–35; Korkiakangas and Passarotti 2011: 108) were not attached to dependency trees, because they were not relevant to the author's research interests at the time. From the drafting of LLCT2 onwards, all sentence adverbials and emphasizing particles are now attached, while punctuation remains omitted. Given that the description of sentence adverbials and emphasizing particles in the *Guidelines* only applies well to Classical Latin, the principles outlined in the *Guidelines* are only followed in marking negating participles with AuxZ and the first coordinator in correlative coordinations, such as *et … et* 'both … and', with AuxY, while all remaining single-word adverbials and particles are marked as ADV (adverb). In addition, AuxY is used to mark the functionally superfluous part in multiword expressions that had become fixed in charter Latin, like *una* in *una cum* 'together with'.

Perhaps the most relevant departure from the *Guidelines* is that not all non-subject complements of a verb are marked as OBJ (object), as suggested by the *Guidelines*, but only direct and indirect objects and the agents of passive verbs (*Guidelines*: 13, Korkiakangas, 2016: 35). The reason for this decision is twofold: first, the valency schemes of verbs are unstable in charter Latin, and it is sometimes impossible to know what was really intended; second, using OBJ for all types of complements in Early Medieval Latin is potentially risky, because its noun morphology does not distinguish between different types of OBJ, as was the case in Classical Latin. As an example of this second concern, OBJ is assigned to prepositional phrases in the *Guidelines* in the case of both *revertitur in ecclesia* 'returns to the church' and *offeruit in ecclesia* 'donated to the church', though the former is a complement of direction and the latter a complement of indirect object, a typical construction of charter Latin where prepositional phrases, in lieu of the Classical dative case, express the Recipient role. The former is annotated as ADV and the latter as OBJ. This practice seems to be sufficient to distinguish direct from indirect objects, because the verbs that govern a Recipient complement and, thus, an indirect object, are limited to a few lexemes that can be easily excluded where needed.

It is not syntax but morphology that requires most new annotation rules in LLCT in comparison to Classical Latin. Korkiakangas and Passarotti (2011) proposes an etymology principle which reduces non-standard word forms to their Classical Latin ancestors, because, on the whole, the scribes still pursued Classical Latin grammar. For example, a non-standard plural accusative form *quam* 'this' in (1) is annotated like the corresponding Classical *quas* and *Warnegausu* like the singular accusative *Warnegausum*. The etymology principle also has implications concerning lemmas: for

[3] PDT: https://ufal.mff.cuni.cz/pdt3.0

example, the form *ad* is actually meant to be *ab* 'from' instead of *ad* 'to' and is thus lemmatized as *ab*. In (1), an LLCT phrase is compared to its equivalent in Classical Latin (in inverted commas), followed by a word-by-word glossa and an approximate English translation.

(1) *quam biro cartolas binditionis nostres ad nus factas Warnegausu not*(*arium*) *iscribere tradedimus*

'*qu-as*	*vero*	*chartul-as*	*venditionis nostrae*	*ab nobis*	*fact-as*
this-ACC.PL	verily	charter-ACC.PL	of our sale	by us	made-ACC.PL

Warnegaus-um	*notari-um*	*scribe-re*	*roga-vi-mus*'
Warnegausus-ACC.SG	notary-ACC.SG	to write	ask-PF-1PL

We asked the notary Warnegausu to write these sales contracts we made.

(*CDL* 66, AD 738)

LLCT also features a diplomatic annotation layer that indicates whether a sentence belongs to the formulaic or non-formulaic part of the charter. As mentioned in section 2, formulaic parts were anchored to an age-old legal tradition, alien to the everyday language, while the non-formulaic parts reflect the spoken idiom. Korkiakangas (2018) explores this diplomatic information, investigating how much the language differs between the formulaic and non-formulaic parts.

Abbreviations are frequent in Early Medieval writing. All the words whose inflectional endings were abbreviated, like *not*(*arium*) in (1), are marked with an *expan* tag in LLCT. However, abbreviated words are lemmatized and tagged morphologically and syntactically like normal, unabbreviated words (see Figure 1). The same applies to fully or partly damaged words, which are restored whenever possible and marked with a *damage* tag. Especially when the expanded and damaged words lack their inflectional endings, they cannot be used in linguistic analysis, but they can figure as nodes to build a proper syntactic tree. This is an advantage of formulaic language, where even damaged passages can be reconstructed with a high degree of reliability. The treatment of abbreviated and damaged words is explained in detail in Korkiakangas and Lassila (2013). The diplomatic mark-up concerning formulaic and non-formulaic parts as well as the abbreviations and restorations is currently available in the PML versions of LLCT1 and LLCT2. Figure 1 presents an extract from the PML version of LLCT1 featuring the quote (1).

Figure 1. Quote (1) as hierarchically nested XML (PML).

```
<LM id="2934875" document_id="CDL" subdoc="66" date="738" place="Massa Mustiba (Chiusi)" scribe="Uuarnegausu"
 type="Charta venditionis" redaction="original">
  <LM id="2" relation="UNDEFINED" segmentation="formulaic" status="normal" form="biro" lemma="vero1"></LM>
  <LM id="12" relation="PRED" segmentation="formulaic" status="normal" form="tradedimus" lemma="trado1" pos=
  "verb" person="first_person" number="plural" tense="perfect" mood="indicative" voice="active">
    <LM id="11" relation="OBJ" segmentation="formulaic" status="normal" form="iscriberes" lemma="scribo1" pos
    ="verb" tense="present" mood="infinitive" voice="active">
      <LM id="3" relation="OBJ" segmentation="formulaic" status="normal" declension="1" animacy="inanimate"
      indirect_object="-" form="cartolas" lemma="chartula1" pos="noun" number="plural" gender="feminine" case
      ="accusative">
        <LM id="1" relation="ATR" segmentation="formulaic" status="normal" declension="1" animacy="inanimate"
         indirect_object="-" form="quam" lemma="qui1" pos="pronoun" number="plural" gender="feminine" case=
        "accusative"></LM>
        <LM id="4" relation="ATR" segmentation="formulaic" status="normal" declension="3" animacy="inanimate"
         indirect_object="-" form="binditionis" lemma="venditio1" pos="noun" number="singular" gender=
        "feminine" case="genitive">
          <LM id="5" relation="ATR" segmentation="formulaic" status="normal" declension="1" animacy=
          "inanimate" indirect_object="-" form="nostres" lemma="noster1" pos="adjective" number="singular"
          gender="feminine" case="genitive"></LM>
        </LM>
        <LM id="8" relation="ATR" segmentation="formulaic" status="normal" declension="1" animacy="inanimate"
         indirect_object="-" form="factas" lemma="facio1" pos="participle" number="plural" tense="perfect"
        mood="participial" voice="passive" gender="feminine" case="accusative">
          <LM id="6" relation="AuxP" segmentation="formulaic" status="normal" form="ad" lemma="ab1" pos=
          "preposition">
            <LM id="7" relation="OBJ" segmentation="formulaic" status="normal" declension="2" animacy=
            "animate" indirect_object="-" form="nus" lemma="nos1" pos="pronoun" number="plural" gender=
            "masculine" case="accusative"></LM>
          </LM>
        </LM>
      </LM>
      <LM id="9" relation="SBJ" segmentation="formulaic" status="normal" declension="2" animacy="personal"
      indirect_object="-" form="Uuarnegausu" lemma="Warnegausus1" pos="noun" number="singular" gender=
      "masculine" case="accusative">
        <LM id="10" relation="ATR" segmentation="formulaic" status="expan" declension="2" animacy="animate"
        indirect_object="-" form="notarium" lemma="notarius1" pos="noun" number="singular" gender="masculine"
         case="accusative"></LM>
```

The following section presents what is specific to each of the three parts of LLCT.

5. Parts of LLCT

5.1. LLCT1

LLCT1 was constructed between 2010 and 2014 for the author's then-ongoing PhD project, which analysed the accusative-subjects of finite verbs. The manual annotation was checked for quality, and the practices that had changed in the process were harmonized. However, in comparison with LLCT2 and LLCT3, the annotation of LLCT1 looks incoherent and should clearly be revised again.

The sources of LLCT1 are three copyright-free editions. *Codice diplomatico longobardo* 1–2 (CDL)[4] is available online as a well-checked web text, while *Codice diplomatico toscano* 2:1 (CDT) and *Memorie e documenti per servire all'istoria del Ducato di Lucca* 5:2 (MED) are digitized by Google Books, so they had to be proofread carefully. As the latter two source editions were completely outdated, their text was checked against the modern *Chartae Latinae Antiquiores* (ChLA) series 1 and 2, which are, however, under copyright and could not therefore be used directly. Since the last 21 MED charters between 865 and 869 were not yet published in ChLA, their readings were checked against the originals at the Archivio storico diocesano in Lucca. This was because the author only wanted to include full decades.

In all, 519 charters were included that were written in Tuscia and that were not too fragmentary. Each decade thus comprises 35–40 charters, except for the first four decades (710s to 740s), which are meagre with documents. Tuscia is a region that corresponds to much of modern Tuscany. No

[4] CDL: https://www.oeaw.ac.at/gema/langobarden/lango_urk.htm

diatopic investigation was planned at the time, and Tuscia is a natural choice, since most of the surviving Early Medieval Italian documents are Tuscan. To ensure that the readings do in fact reflect the Early Medieval linguistic situation, only documents that were originals or coeval copies were admitted, not documents that survive as Late Medieval or Early Modern copies.

As stated in section 4, non-nominal sentence adverbials and emphasizing particles were excluded from the dependency trees of LLCT1. Another defect is that MED, one of the base editions of LLCT1, usually truncates the subscriptions, which always follow a strict formula: e.g., *ego Andreas rogatus etc*. instead of *ego Andreas rogatus ab Aloni me teste subscripsi* 'I, Andreas, asked by Aloni, signed as witness'. Although they are not particularly useful linguistically due to their extremely formulaic nature, the 688 truncated subscriptions of LLCT1 should be completed with the help of ChLA, when the annotation of LLCT1 is revised anew. For the present, the most up-to-date version of LLCT1 is available at Zenodo in Prague Markup Language (PML) format.[5]

5.2. LLCT2

LLCT2 was built between 2016 and 2018 as a chronological extension of LLCT1. All 521 charters of LLCT2 are Tuscan, and they extend the time range to 897. The charters of the decades prior to the 870s, which had been omitted earlier in LLCT1, were also added. Thus, LLCT1 and LLCT2 together contain almost all the available non-fragmentary Tuscan original charters and coeval copies until 900.

The LLCT2 charters derive from four 19th-century editions: MED, MED2, MED3 and MED4. They were again carefully corrected using ChLA as benchmark, but the text is still essentially that of the original editions, given that all the abbreviation marks and line breaks present in ChLA are lacking. LLCT1 served as the training data for the annotation. The proofread text was lemmatized with a simple multi-replace script that matched similar strings to their LLCT1 lemmas. Since the lexicon is repetitive, this resulted in a retrieval of 97% and an accuracy of almost 100%. TnT Tagger was used to tag the morphology, with 88% accuracy. The syntactic dependencies were parsed by training MaltParser on LLCT1.[6] The lemmas and morphological tags were subsequently corrected manually in Excel and the syntactic dependencies and tags in the Arethusa annotation environment. Finally, textual annotation regarding formulaic and non-formulaic parts as well as abbreviations and damaged words was added manually. Manual correction pursued a zero-error rate, so the accuracy of the annotation in its present state approaches 100%.

In 2019, the LLCT2 annotation was thoroughly revised and converted to Universal Dependencies in a joint effort with the Linking Latin (LiLa) project of the University of Sacred Heart in Milan. The CoNLL-U format data will be distributed in a subsequent release of the Universal Dependencies at the project's website.[7] Until then, a CoNLL version of LLCT2 can be downloaded at Zenodo.[8]

5.3. LLCT3

Once completed in 2021, LLCT3 will be the smallest of the LLCT treebanks. It will consist of three sets of data, which extend both the chronological span within Tuscia and, for the first time, will expand outside of Tuscia.

[5] https://zenodo.org/record/1197357#.XbmDlmZS9EY
[6] TnT Tagger: http://www.coli.uni-saarland.de/~thorsten/tnt/, MaltParser: http://www.maltparser.org
[7] https://universaldependencies.org/#language-
[8] https://zenodo.org/record/3522868#.XbmCo2ZS9EY

As for Tuscia, LLCT3 will be based on the author's new diplomatic edition of 72 tenth-century charters preserved in the Archivio storico diocesano of Lucca (AD 900–1000, c. 48,800 tokens). The other source-texts are available in electronic format: 102 southern Italian charters from Campania from AD 792–899 in the online version of the *Codice diplomatico cavense* (CDC, c. 42,500 tokens)[9], 44 northern Italian charters mainly from Lombardy and Emilia from AD 721–799 and three previously omitted Tuscan charters from AD 732–768 in the above-mentioned online version of CDL (c. 19,100 tokens). These editions' readings will be again checked against ChLA and annotated using LLCT2 as the training data. In total, LLCT3 will comprise 221 charters and around 110,400 tokens.

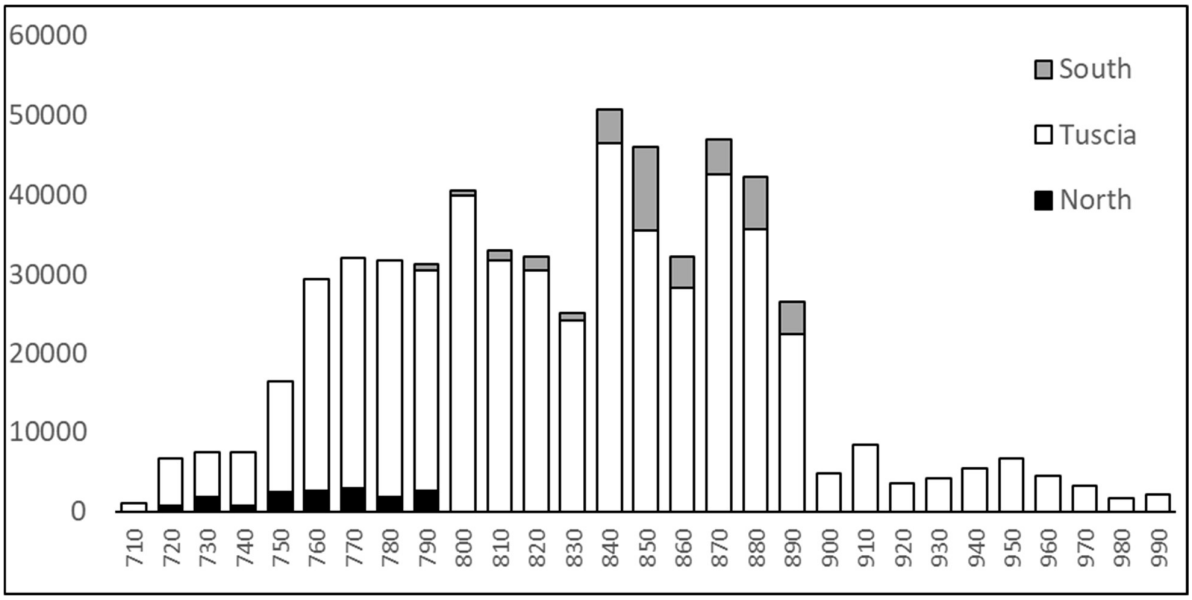
6. Geographical and chronological coverage

The geographical and chronological distribution of the LLCT charters determines the research questions which can be asked. Historical corpus linguistics would benefit from contrasting the Early Medieval Latin of northern Italy with that of central and southern Italy, because the border between the later western and eastern Romance languages arguably runs across northern and central Italy (Bossong, 2016: 67). This so-called La Spezia–Rimini line is based mostly on phonological differences—hence the need for a treebank-based syntactic examination.

LLCT1 and LLCT2 are restricted to Tuscia, which is traditionally considered part of the Eastern Romance. Nevertheless, diatopic microvariation in local scribal conventions and traditions can be examined even within Tuscia (e.g. Korkiakangas and Lassila, 2018). Proper diatopic comparisons will be possible with the completion of LLCT3, though, unfortunately, the northern charters are few, and the southern ones relatively late. That Tuscia dominates the data, with 91% of all LLCT tokens, reflects how unevenly documents have survived. On the other hand, there are more northern documents from after 800, but they cannot be used, because they are under copyright. As Figure 2 shows, the largest challenge for diatopic study is that the corpus of northern charters does not overlap chronologically with the southern charters. To ensure that geographical variation is not chronological, the geographically contrasting sets should be contemporary.

Figure 2. Geographical and chronological coverage of LLCT1, LLCT2 and LLCT3 (tokens by decade).

[9] CDC at the Archivio della Latinità Italiana del Medioevo (ALIM): http://www.alim.dfll.univr.it/Notarili/alimnot.nsf/RPD/811D99C520B01FE6C12571E900558A31!opendocument&vs=RPD



The corpus is also not chronologically balanced, but this is seldom the case with historical corpora. As noted earlier, the first 50 years are scant. The bars reflect quite accurately all the sources that have survived to us until 900. After that, the remaining bars indicate the charters transcribed in Lucca for LLCT3.

7. Conclusion

The three parts of LLCT together constitute a substantial corpus of non-standard non-literary Latin that reflects the language change that had taken place in the Early Medieval Latin of Italy. LLCT1 was mostly annotated manually, while LLCT2 was annotated by training a parser on LLCT1 (and LLCT3 will be annotated on LLCT2), followed by manual correction. As LLCT3 will expand the geographical scope outside Tuscia, diatopic analyses will be viable, though the resulting geographical distribution is chronologically highly biased. The textual annotation layer makes it possible to restrict searches to the formulaic or non-formulaic parts of charters and to words that are not damaged or abbreviated.

LLCT1 and LLCT2 are currently open access, as will be the case with LLCT3 once finished in 2021. In the future, the plan is to merge LLCT1, LLCT2, and LLCT3 into one unified treebank, with a coherent annotation and metadata structure, preferably in both the LDT and Universal Dependencies formats.

References

Bamman, D., M. Passarotti, G. Crane and S. Raynaud. 2007. *Guidelines for the syntactic annotation of Latin treebanks* (v. 1.3).
(https://itreebank.marginalia.it/doc/2007_Passa+Bamman+Crane+Raynaud_Guidelines%20Tb.pdf)

Bossong, G. 2016. ‘Classification’, in A. Ledgeway and M. Maiden (eds.) *The Oxford Guide to the Romance Languages*, pp. 63–72. Oxford: Oxford University Press.

CDC = *Codice diplomatico cavense* 1. Ed. by M. Schiani, M. Morcaldi and S. De Stefano. Napoli: Piazzi, 1873

CDL = *Codice diplomatico longobardo* 1–2. Ed. by L. Schiaparelli. Roma: Tipografia del Senato, 1929–1933.

CDT = *Codice diplomatico toscano* 2:1. Ed. by F. Brunetti. Firenze: Leopoldo Allegrini e Giov. Mazzoni, 1833.

ChLA1 = *Chartae Latinae Antiquiores*. Facsimile-edition of the Latin Charters Prior to the Ninth Century. Ed. by A. Bruckner, R. Marichal and al. Olten, Dietikon, Zürich: Urs Graf Verlag, 1954–2001.

ChLA2 = *Chartae Latinae Antiquiores*. Facsimile-edition of the Latin Charters. 2nd Series: Ninth Century. Ed. by G. Cavallo, G. Nicolaj and al. Dietikon, Zürich: Urs Graf Verlag, 1997–2019.

Frank-Job, B. and M. Selig. 2016. ‘Early evidence and sources’, in A. Ledgeway and M. Maiden (eds.) *The Oxford Guide to the Romance Languages*, pp. 24–34. Oxford: Oxford University Press.

Hajič, J., J. Panevová, E. Buráňová, Z. Urešová and A. Bémová. 1999. *Annotations at Analytical Level*. Instructions for annotators.
http://ufal.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/aman_en.pdf

McGillivray, B. 2014. *Methods in Latin Computational Linguistics*. Leiden, Boston: Brill.

MED = *Memorie e documenti per servire all'istoria del Ducato di Lucca* 5:2. Ed. by D. Barsocchini. Lucca: Francesco Bertini, 1837.

MED2 = *Memorie e documenti per servire all'istoria del Ducato di Lucca* 5:3. Ed. by D. Barsocchini. Lucca: Francesco Bertini, 1841.

MED3 = *Memorie e documenti per servire all'istoria del Ducato di Lucca* 4:1. Ed. by D. Bertini. Lucca: Francesco Bertini, 1818.

MED4 = *Memorie e documenti per servire all'istoria del Ducato di Lucca* 4:2. Ed. by D. Bertini. Lucca: Francesco Bertini, 1836.

Pinkster, H. 1990. *Latin Syntax and Semantics*. London: Routledge.

Pratesi, A. 1979. *Genesi e forme del documento medievale*. Roma: Jouvence.

Korkiakangas, T. 2016. *Subject case in the Latin of Tuscan charters of the 8th and 9th centuries*. Helsinki: Societas Scientiarum Fennica.

Korkiakangas, T. 2018. ‘Spoken Latin Behind Written Texts: Formulaicity and Salience in Medieval Documentary Texts’, *Diachronica* 35, pp. 429–449.

Korkiakangas, T. and M. Lassila. 2013. ‘Abbreviations, fragmentary words, formulaic language: Treebanking medieval charter material’, in F. Mambrini, M. Passarotti and C. Sporleder (eds.) *Proceedings of the third workshop on annotation of corpora for research in the humanities*, pp. 61–72. Sofia: Bulgarian Academy of Sciences.

Korkiakangas, T. and M. Lassila. 2018. ‘Visualizing linguistic variation in a network of Latin documents and scribes’, *Journal of Data Mining & Digital Humanities*, art. 4472 (https://jdmdh.episciences.org/4472)

Korkiakangas, T. and M. Passarotti. 2011. ‘Challenges in annotating Medieval Latin charters’, *Journal of Language Technology and Computational Linguistics* 26, pp. 103–114.

Sabatini, F. 1965. ‘Esigenze di realismo e dislocazione morfologica in testi preromanzi’, *Rivista di Cultura Classica e Medievale* 7, pp. 972–998.

Sornicola, R. 2017. ‘“Transizione” e “transizioni” dal latino al romanzo: il progetto di analisi linguistica dei documenti cavensi del IX secolo’, in R. Sornicola, E. D’Argenio and P. Greco (eds.) *Sistemi, norme, scritture: La lingua delle più antiche carte cavensi*, pp. 13–25. Napoli: Giannini.

Wright, R. 2016. ‘Latin and Romance in the medieval period’, in A. Ledgeway and M. Maiden (eds.) *The Oxford Guide to the Romance Languages*, pp. 14–23. Oxford: Oxford University Press.